Data S2 - Online supporting information for *Occupancy Modeling Species-Environment Relationships with Non-ignorable Sampling Designs*

Kathryn M. Irvine, Thomas J. Rodhouse, Wilson J. Wright, Anthony R. Olsen

Data S2: R-code for simulation investigation into impacts of design specifications on occupancy parameter estimates using pseudo- and traditional maximum likelihood.

This Appendix includes R scripts with functions for conducting simulations using sample weights and occupancy models with user specified covariate data. We also provide a demonstration using the example from the paper. These files require the dplyr and ggplot2 packages.

Files Included:

1. pmle_sim_functions_new.R:

   creates functions for performing simulation studies. This file includes six functions in total. Each has comments describing the inputs and how they are used. We also include brief descriptions of every function here.

   (a) *logL.fun*: calculates the log-likelihood of an occupancy model given supplied model matrices for the occupancy and detection model components. Can include an argument for weights to perform pseudo-maximum likelihood estimation.

(b) *occ_pmle*: fits a model based on supplied detection and occupancy formulas, detection history matrix, site-level covariates, visit-level covariates, and (optional) sample weights. If no sample weights are given then standard maximum-likelihood estimation is performed. With sample weights, then pseudo-maximum likelihood estimates are obtained.

(c) *sim.fun*: simulation wrapper that generates population data based on a supplied generating model, coefficient values, and covariate data. For each iteration a new population is simulated. For every population, a sample is taken following each of the given designs and all of the supplied models are fit to each dataset. Designs are specified by including vectors for strata identifiers (for the entire population defined in the sampling frame) and the corresponding sample sizes for each stratum. A "legacy" strata could be included in the simulation by having the corresponding sample size exactly equal to the number of legacy sites. For generating data and fitting models, this function assumes constant detection probabilities. The output includes the estimates and standard errors for every model fit under the different iterations and designs. It also includes estimates from fitting each model to the census data.

(d) *summary.fun*: helper function used to summarize the output from a simulation.

(e) *sum.results*: summarizes the results from a simulation for a particular parameter of interest from a specified model. User chooses which of the supplied designs to include in this summary. For each supplied design of interest, this function calculates the average 95% CIs (averages of the upper and lower bounds) and average of the point estimates ($\hat{\beta}_{D|M_k}$) across all simulated datasets. The coverage properties for both the weighted (PMLE) and unweighted (MLE) model fits are provided.

(f) *res.plot*: used on the results of a *sum.results* call, creates a plot summarizing the results of interest

2. run_pmle_sims_new.R:

using the example from the paper, shows how to use the functions in the pmle_sim_functions_new.R file to conduct a simulation study. This file includes comments about the output from the various functions and how to use them. The example includes very few iterations in order to save computation time.